DOCUMENT RESUME

ED 395 024 TM 025 041

AUTHOR	Way, Walter D.; Reese, Clyde M.
TITLE	An Investigation of the Use of Simplified IRT Models for Scaling and Equating the TOEFL Test. TOEFL Technical Report TR-2.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-TR-90-29
PUB DATE	Feb 91
NOTE	41p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Correlation; *Equated Scores; *Estimation (Mathematics); Goodness of Fit; *Item Response Theory; Models; *Raw Scores; Sample Size; *Scaling; Scores; Simulation
IDENTIFIERS	Item Parameters; One Parameter Model; *Test of English as a Foreign Language; *Three Parameter Model; Two Parameter Model

ABSTRACT

The use of two alternative item response theory (IRT) estimation models in the scaling and equating of the Test of English as a Foreign Language (TOEFL) was explored; and item scaling and test equating results based on these models were compared with results based on the three-parameter (3PL) model currently being used with the TOEFL. Models were a modified one-parameter (M1PL) and a modified two-parameter (M2PL). Simulated equatings were compared in terms of correlations between estimated and generating parameters, model-data fit, and concordance of simulated score conversions, with conversions based on the generating parameters. Results clearly indicated that the 3PL model performed better than the M1PL and M2PL models. There was also evidence that the M2PL model performed better than the M1PL. Discrepancies between score conversions based on the M1PL and the M2PL models and those based on the 3PL model tended to occur at the lower and upper ends of the score scales. Results of the study for the 3PL model indicated that while correlations between item parameter estimates and generating parameters tended to be affected by sample size, neither the quality of model-data fit nor the quality of simulated equatings was sensitive to sample size. Three appendixes present the raw-to-scaled score conversions for three TOEFL sections. (Contains 3 figures, 8 tables, 12 appendix tables, and 16 references.) (Author/SLD)

TOEFL®

February 1991

Technical Report

TR- 2

An Investigation of the Use of Simplified IRT Models for Scaling and Equating the TOEFL Test

By Walter D. Way and
Clyde M. Reese

An Investigation of the Use of Simplified IRT Models for Scaling and Equating the TOEFL Test

by

Walter D. Way

Clyde M. Reese

Educational Testing Service
Princeton, New Jersey 08541

TR-90-29

Acknowledgments

The authors thank Patricia Carey and Alison Snieckus for programming support, and Marna Golub-Smith, Nancy Petersen, and Martha Stocking for reviewing earlier drafts of this manuscript.

The Test of English as a Foreign Language (TOEFL) was developed in 1963 by a National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of more than thirty organizations, public and private, that were concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program, and in 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and the GRE Board and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

✣ ✣ ✣

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English as a second language specialists from the academic community. Currently the Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. However, many projects require the cooperation of other institutions, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. Local research may sometimes require access to TOEFL data. In such cases, the program may provide the data following approval by the Research Committee. All TOEFL research projects must undergo appropriate ETS review to ascertain that the confidentiality of data will be protected.

Current (1990-91) members of the TOEFL Research Committee are:

Patricia L. Carrell (Chair)	University of Akron
James Dean Brown	University of Hawaii
Patricia Dunkel	Pennsylvania State University
Fred Genesee	McGill University
Elliott Judd	University of Illinois at Chicago
Elizabeth C. Traugott	Stanford University

Abstract

The purpose of this study was to explore the use of two alternative item response theory estimation models in the scaling and equating of TOEFL -- a modified one-parameter model (M1PL) and a modified two-parameter model (M2PL) -- and to compare item scaling and test equating results based on these two alternative models with results based on the three-parameter model (3PL) that is currently being used to scale and equate the TOEFL. The study employed a design in which a typical TOEFL equating was simulated using artificial data. The simulated equatings were compared in terms of correlations between estimated and generating parameters, model-data fit, and concordance of simulated score conversions with conversions based on the generating parameters.

The results of the study clearly indicated that the 3PL model performed better than the M1PL and M2PL models on the basis of each of the evaluation criteria. There was also evidence that the M2PL model performed better than the M1PL model, particularly in terms of model-data fit and in the weighted root mean square difference statistics used to evaluate the simulated score conversions. The results of the study also indicated that discrepancies between score conversions based on the M1PL and M2PL model and those based on the 3PL model tended to occur at the lower and upper ends of the score scales. Finally, the results of the study for the 3PL model indicated that while correlations between item parameter estimates and generating parameters tended to be affected by sample size, neither the quality of model-data fit nor the quality of simulated equatings appeared to be sensitive to the different sample sizes used in the study.

Table of Contents

List of Tables

List of Figures

Background

The present investigation examines the use of three logistic models for scaling and equating the Test of English as a Foreign Language (TOEFL®). TOEFL is a multiple-choice test designed to assess the English language proficiency of foreign students wishing to gain admission to institutions of higher learning in the United States or Canada. The test consists of three separately timed sections-- Listening Comprehension (Section 1), Structure and Written Expression (Section 2), and Reading Comprehension and Vocabulary (Section 3). Each section is scored and equated separately. Section raw scores are computed as the number of correct responses and are converted to the TOEFL scale using item response theory (IRT) true score equating (Lord, 1980).

Since 1978, the three-parameter logistic (3PL) model has been used to scale and equate the TOEFL test. In this model, the probability of a correct response (P_i) is a function of the examinee's ability and three item parameters and can be stated by:

$$\text{3PL:}\quad P_i(\theta_j) = c_i + (1-c_i)/\{1+\exp[-Da_i(\theta_j-b_i)]\}, \qquad (1)$$

where c_i is the pseudo-guessing parameter for item i, a_i is the item discrimination parameter for item i, b_i is the item difficulty parameter for item i, θ_j is the ability of examinee j, and D is a constant assuming the value of 1.7 (which is employed to make the logistic curve closely approximate the normal ogive model).

The two more restrictive models considered in the present investigation are (a) a modified two-parameter logistic (M2PL) and (b) a modified one-parameter logistic (M1PL). The M2PL model assumes a constant, nonzero value for the pseudo-guessing parameter (i.e., $c_i = c > 0$ for all i). The M1PL also assumes a constant, nonzero value for c and makes the additional assumption that the discrimination parameters for each item are equal (i.e., $a_i = a$ for all i). The M2PL and M1PL models can be stated by:

$$\text{M2PL:}\quad P_i(\theta_j) = c + (1-c)/\{1+\exp[-Da_i(\theta_j-b_i)]\}, \qquad (2)$$

$$\text{M1PL:}\quad P_i(\theta_j) = c + (1-c)/\{1+\exp[-Da(\theta_j-b_i)]\}, \qquad (3)$$

where c is a constant value based on a priori assumptions (e.g., guessing is random, therefore c is equal to the reciprocal of the number of alternatives) or empirical evidence (e.g., the average of the c_i estimates for the items from previous 3PL calibrations) and a is a constant value of the item discrimination parameter for all items. In the present study, c was fixed for each section at a constant value based on the median of the pseudo-guessing parameter estimates over several previous TOEFL administrations.

The choice of an IRT model is usually based on a priori expectations and empirical investigations of model-fit (Hambleton & Swaminathan, 1985). In the case of TOEFL, the multiple-choice nature of the test implies that examinee guessing does exist, suggesting the need to fit a 3PL model. However, in a previous study with the TOEFL test, Hicks (1984) found that a M1PL model did produce reasonable equating results, even though it was clear that application of the 1PL model would require acceptance of some inadequately fit items. If the use of one of the more restrictive models was found to be feasible for scaling and equating TOEFL, several benefits would result. For example, the sample sizes needed for IRT pretest item calibrations would be substantially reduced, as would associated computer costs. This study explores the use of two alternative IRT models for equating TOEFL, and compares them to the 3PL model which is presently being used.

Description of TOEFL Equating Design

Originally, IRT equating of TOEFL was carried out using a method called fixed b's scaling (Hicks, 1983). Beginning in January 1989, the equating procedures for TOEFL changed so that each of the three test sections is scaled using an external anchor in equating/pretesting administrations. In administrations using both equating and pretest items, examinees typically take one of four versions, or scramble forms. For all scramble forms, examinees take the same operational items for each section (although for Sections 2 and 3 the items in the test booklets are in different orders). However, one of the scramble forms contains a set of external equating items for each section, while each of the other three scramble forms contain a unique set of pretest items. The total number of items given in a TOEFL pretest administration can be broken down as follows:

	Section I					Section II					Section III				
Scramble Form	Op.	Eq.	P1	P2	P3	Op.	Eq.	P1	P2	P3	Op.	Eq.	P1	P2	P3
A	50	30	--	--	--	38	20	--	--	--	58	30	--	--	--
B	50	--	30	--	--	38	--	20	--	--	58	--	30	--	--
C	50	--	--	30	--	38	--	--	20	--	58	--	--	30	--
D	50	--	--	--	30	38	--	--	--	20	58	--	--	--	30

Op. = Operational items, Eq. = Equating items, P1, P2, P3 = Pretest items.

Responses to a total of 466 items (170 Section 1, 118 Section 2, and 178 Section 3 items) are collected in a typical TOEFL pretest administration. For each section, there are four sets of external equating items that are used in rotation for the equating/pretesting administrations. It should be noted that under this equating design, rather than scaling new form item parameter estimates to a base form through previous forms, scaling is carried out through the external equating items directly to the base form scale. Items are calibrated using the 3PL model and transformed to the scale of the base form by finding the scaling parameters that relate the difficulty and discrimination estimates for the external equating items to their original estimates. These transformation parameters are then applied to the difficulty and discrimination estimates for all items in the administration. In all cases, items are calibrated using the computer program LOGIST (Wingersky, Barton, & Lord, 1982; Wingersky, Patrick, & Lord, 1988) and transformed using an item characteristic curve (ICC) method developed by Stocking and Lord (1983). Finally, scores on the operational items for each section are equated to the base form using IRT true score equating.

Is the 3PL Model Necessary with TOEFL?

One way to reduce the sample sizes necessary for successful equating would be to use a M2PL or M1PL model with TOEFL data. Despite the a priori assumption that guessing exists on multiple-choice tests, these models may produce equally stable and accurate item parameter estimates and equating conversions because estimation of the pseudo-guessing parameter is fraught with difficulties. It is well known from the IRT literature that the estimation of the pseudo-guessing parameter is often problematic with easier items due to lack of information at the low end of the ability scale (Baker, 1987; Lord, 1980). In a review of the IRT item parameter estimation literature, Baker (1987) concluded that the pseudo-guessing parameter is poorly estimated in terms of both bias and standard error, and that c-estimates do not correlate well with the underlying values in simulation studies. Furthermore, he pointed out that estimation problems with the c-parameter do not occur in isolation:

> Troubles in the estimation of c tend to carry over into the estimation of b and a. In particular, the estimation of difficulty is affected as error in c results in a shift in b. The carryover is also evident in the larger standard errors of a and b. (p. 134)

Because of the problems associated with estimating c, LOGIST assigns equal minimum c-values to items with ill-defined parameter estimates. It employs a stability criterion for each item, b - 2/a, defined as the ability level at which the proportion of correct responses is only about .03 above the value of c (Wingersky, Barton, & Lord, 1982). All items with values of b - 2/a less than a specified minimum are automatically assigned c-values equal to a single common value.

Despite the b - 2/a criterion and a number of other constraints designed to stabilize estimation with LOGIST, the possibility that poorly estimated c-values will distort estimates of item difficulty and item discrimination still exists. Poorly estimated c-values are particularly likely with TOEFL because most examinees are at ability levels above the typical item difficulty levels, and only abilities well below item difficulty are informative about c (Stocking, 1988). Thus, it appears reasonable to consider the investigation of alternative IRT models with TOEFL that do not require the estimation of the pseudo-guessing parameter.

A number of studies have compared the 3PL model with 2PL and/or 1PL models in terms of model-data fit (Hambleton & Murray, 1983; McKinley & Mills, 1985; Swaminathan & Gifford, 1979; Yen, 1981). In general, results of these studies suggest that the 3PL model will provide better fit at the item level than the 2PL or 1PL models, unless the data are simulated to fit these latter models. McKinley and Mills (1985) reported that when data were generated with the 3PL model, the 2PL model showed considerably more misfit than the 3PL model in terms of the proportions of items identified as misfitting by several goodness-of-fit statistics. However, under similar conditions, Yen (1981) found that the 2PL model fit the data almost as well as the 3PL model. In both of these studies, the 1PL model provided poor fit except when data were simulated to fit the 1PL model. Previous studies by Marco, Wingersky, and Douglass (1985) and Hicks (1984) suggested that a M1PL model produced equating results that compared favorably with equating results based on the 3PL model. However, it should be noted that these studies employed designs where a test was equated to itself through a series of "chains." Such a design is less relevent to the current TOEFL in that equating is carried out directly to the scale of the base form rather than through a chain of previously administered forms.

Many other studies have compared various IRT models in equating situations, however the breadth and complexities of the designs employed with various IRT equating research have made it difficult to point to definitive conclusions. Skaggs and Lissitz (1986) concluded a review of the IRT equating literature by stressing that the results of various IRT equating methods seem to depend greatly on the context in which the equating occurs. In the case of TOEFL, the equating design represents a unique situation in that external anchor sets of items are employed and scaling to the base form is done directly rather than through chaining. Thus, any judgments about the applicability of alternative IRT models to the TOEFL program would seem to requ're research that employs the TOEFL equating design.

Method

Generation of the Simulated Data

In order to generate realistic simulated TOEFL data, 3PL model item parameter estimates from a recent TOEFL administration transformed to the IRT scale of the TOEFL base form were used as the generating parameters for the simulated data. In addition, systematic samples of ability estimates were taken from the complete set of TOEFL ability estimates for the same TOEFL administration and were used as the generating ability parameters for the simulated data.

It should be pointed out that the simulated data for the study were generated to fit the 3PL model, and the base form item parameters used for all true score equatings were also based on the 3PL model. There were two reasons for this. First, it is fairly certain that guessing does occur on TOEFL, and the most severe test of the adequacy of the more restrictive IRT models will occur when they are applied to data that incorporate guessing. Second, both the TOEFL item bank and the existing TOEFL scale are based on the 3PL model. In order for an alternative item response model to be used with TOEFL it will be necessary to link into the existing 3PL scale, and to utilize "old" item parameter estimates that are based on the 3PL model. In short, because the operational psychometrics of the TOEFL test are 3PL-based, it was believed appropriate to base data simulation on the 3PL model. Based on the generating model, it could be expected that the 3PL model would produce better results than the M2PL and M1PL model in terms of bias. However, no companion predictions could be made with respect to estimation error (M. Stocking, personal communication, February 25, 1990).

For each TOEFL test section, simulated data sets were generated as follows: for each simulated item i and simulee j, a 0 or 1 response was assigned by comparing the probability of correct response as indicated by the 3PL model with the item i and person j parameter values to a random number drawn from a uniform 0, 1 distribution. If the probability of correct response exceeded the value of the uniform random number, the item was scored as correct; otherwise, the item was scored as incorrect. Simulated response strings included responses to items in all three operational sections. For approximately one-fourth of the total simulated data, responses to the equating set items were also generated. Responses to the equating set items for the other three-fourths of the simulated examinees were set to missing. Data were simulated with four different sample sizes, so that the numbers of simulated examinees taking the equating set items were 600, 900, 1200, and 1500. The total sample sizes for the simulated data sets were 2,400, 3,600, 4,800, and 6,000. For all models and sample sizes, comparisons were made using simulated data for each of the three sections of the TOEFL.

Equating the Simulated Data

The simulated data were equated in three steps. In the first step, the simulated data were calibrated using each of the three IRT models investigated in the study (3PL, M2PL, and M1PL). For all calibrations, the LOGIST program was used. For the relevant models, the pseudo-guessing parameter was fixed at a constant value based on the median of pseudo-guessing parameter estimates over several previous administrations of TOEFL. In the second step, the item and ability parameter estimates obtained from each calibration were transformed to the scale of the base form item parameters using constants obtained by applying the ICC transformation methods to the item parameter estimates for the anchor items. In each case, one set of estimates for the anchor items was the 3PL estimates that are used to carry out ICC transformations in operational TOEFL equatings, and the other set was the estimates resulting from applying the given model to the simulated data.[1]

In the third step, TOEFL number-right true scores for each calibration were equated to number-right true scores on the base form using IRT true score equating. The result was an unrounded scaled score on the base form scale for each number-right true score for each calibration. In addition, true score equating was carried out using the generating item parameters for the simulated data. The conversions based on these "true" parameters were used as the criterion for comparing the conversions based on the experimental calibrations.

Evaluating the Results

The results of the study were evaluated in three ways: an examination of relationships between transformed estimates and generating parameters, model-data fit, and correspondence of the equating conversions with the criterion. To assess model-data fit, plots of item-ability regressions and standardized residuals were examined. Particular attention was given to the lower ability intervals, as it is in these intervals where error due to the effects of guessing is likely to occur. The item-ability regression plots graphically present the actual and predicted proportions correct as a function of ability or theta. The standardized residual analyses also examine actual and predicted proportions correct but in a slightly different manner. For the present study, ability or theta estimates for the simulated sample were first grouped into deciles. Then, within each of the decile groups, the residual (actual-predicted) proportion correct was found and standardized using the standard deviation [$\sqrt{P_A(1 - P_A)}$, where P_A is the actual proportion correct] of the responses for that decile group. To summarize the results of the residual analyses, percentages of standardized residuals having absolute values between 0 and 1, between 1 and 2, between 2 and 3, and greater than 3 were tabulated for each model, sample size, and section.

[1] It should be pointed out that in order to simulate TOEFL equatings as realistically as possible, the generating item parameters for the equating set items were not the same as the equating set parameters used in the ICC transformations. In other words, in the actual TOEFL administration from which the generating parameters were taken, there were two sets of estimates available for the equating set items: "old" and "new." The "old" set was used as the equating item parameters for the ICC transformation and the "new" set was used for the generation of the simulated data.

The equating conversions were compared to the criterion using a weighted root mean square difference (WRMSD) statistic. This statistic was based on a similar statistic used by Petersen, Marco, and Stewart (1982). The WRMSD statistic is calculated as follows:

$$\text{WRMSD} = [\Sigma f_j d_j^2/n = \Sigma f_j(d_j-\bar{d})^2/n + \bar{d}^2]^{1/2}, \tag{4}$$

and can be broken up into two subcomponents:

$$\text{SD Diff} = [\Sigma f_j(d_j-\bar{d})^2/n]^{1/2}, \tag{5}$$

$$\text{BIAS} = \bar{d}, \tag{6}$$

where $d_j = (t_j'-t_j)$, t_j' is the estimated converted score for raw score x_j, t_j is the criterion converted score for x_j, $\bar{d} = \Sigma f_j d_j/n$, f_j is the frequency of x_j, $n = \Sigma f_j$, and the summation is over the entire range of x (i.e., the raw scores for a given simulated data set).

Rounded raw to scaled score conversion tables were also obtained using each of the three models and compared to the rounded conversions based on the generating item parameters.

Results

Simulated Data Sets

The raw score means and standard deviations for the simulated data sets are presented in Table 1. These values are similar across samples for a given section, and can be considered typical of raw score means and standard deviations obtained in actual TOEFL administrations.

Table 1
Raw Score Means and Standard Deviations for Simulated Data[a]

	Section 1		Section 2		Section 3	
Sample/Item Set	Mean	SD	Mean	SD	Mean	SD
Sample 1						
Operational (N = 2400)	35.84	8.77	27.71	6.94	38.24	11.02
Equating Set (N = 600)	20.99	5.65	14.11	3.95	20.65	5.59
Sample 2						
Operational (N = 3600)	35.81	8.80	27.83	6.91	38.57	10.86
Equating Set (N = 900)	21.04	5.40	14.02	3.94	20.80	5.59
Sample 3						
Operational (N = 4800)	35.91	8.74	27.76	7.00	38.43	11.00
Equating Set (N = 1200)	21.20	5.32	14.08	3.99	20.77	5.37
Sample 4						
Operational (N = 6000)	35.90	8.76	27.80	6.94	38.47	10.95
Equating Set (N = 1500)	21.05	5.32	13.87	4.06	20.68	5.49

[a] Number of operational items per section: 50 (1), 38 (2), 58 (3). Number of equating set items per section: 30 (1), 20 (2), 30 (3).

Relationships between Generating Parameters and Estimates

The relationships between the generating parameters and the transformed item parameter estimates are summarized by section in Tables 2 through 4. The data in these tables include intercorrelations as well as the means and standard deviations of the parameters and corresponding estimates. In comparing mean item discrimination parameter estimates, it can be seen that there is larger variation in the 3PL means across samples compared to the M2PL means, and that the mean item discrimination estimates based on these two models within a given sample are often discrepant. In addition, there is some tendency for the mean 3PL item discrimination estimates to be higher than the corresponding mean item discrimination parameters. In all samples, the transformed constant item discrimination parameter value for the M1PL model is lower than the mean of the generating item discrimination parameters.

A pattern in Tables 2 through 4 that was expected is that the highest correlations between generating item discrimination and item difficulty parameters and corresponding estimates are nearly always those based on 3PL estimates. In the case of item discrimination, the differences between the correlations based on the 3PL estimates and those based on the M2PL estimates are fairly substantial. Correlations between the 3PL item difficulty estimates and the generating parameters tend to be only slightly higher than those for the M2PL and M1PL item difficulty estimates (but note that all correlations are at least .92). Another expected trend borne out by the data in Tables 2 through 4 was that correlations between the 3PL estimates and the generating parameters would be higher in the larger samples. This is seen to be particularly true in comparing the correlations for the operational items with those for the equating set items. In contrast, correlations between the M2PL and M1PL estimates and the generating parameters in the larger samples tend to indicate little or no increase compared to smaller samples.

Data related to the setting of the common c-value by LOGIST for the 3PL estimates are given in Table 5. These data indicate that the common c-value differed from sample to sample in Sections 2 and 3, and that the number of items set at the common c-value did not appear to be related to sample size.

Relationships between the generating ability parameters and the transformed ability estimates based on the three models are summarized in Table 6. These data are extremely consistent across estimation model and sample. In all cases, the M1PL, M2PL, and 3PL ability estimates have nearly identical correlations with the generating ability parameters. Correlations among the M1PL, M2PL, and 3PL ability estimates are higher than any of the correlations between the ability estimates and the generating parameters, and in many cases are nearly perfect. Means of the ability parameter estimates are generally similar to the means of the generating abilities, with largest absolute difference equal to .06 (between Section 2 M2PL estimates and parameters for Sample 1).

Table 2
Summary of Relationships between Generating Item Parameters (Parms) and Estimates (M1PL, M2PL, and 3PL) - Section 1[a]

Data/ Model	Operational Items (n = 50)											
	Item Discrimination				Item Difficulty					Guessing		
	Mean	SD	M2PL	3PL	Mean	SD	M1PL	M2PL	3PL	Mean	SD	3PL
Sample 1 (N = 2400)												
M1PL	1.29	--			-0.16	0.58				0.19	--	
M2PL	1.31	0.49			-0.17	0.62	98			0.19	--	
3PL	1.47	0.51	78		-0.11	0.62	94	95		0.21	0.12	
Parms	1.44	0.45	75	93	-0.09	0.61	93	92	95	0.23	0.14	75
Sample 2 (N=3600)												
M1PL	1.30	--			-0.16	0.59				0.19	--	
M2PL	1.32	0.50			-0.18	0.63	98			0.19	--	
3PL	1.43	0.49	76		-0.12	0.68	94	95		0.22	0.13	
Parms	1.44	0.45	78	97	-0.09	0.61	93	94	99	0.23	0.14	95
Sample 3 (N = 4800)												
M1PL	1.28	--			-0.15	0.59				0.19	--	
M2PL	1.34	0.50			-0.16	0.61	98			0.19	--	
3PL	1.50	0.52	68		-0.08	0.63	93	93		0.22	0.15	
Parms	1.44	0.45	79	96	-0.09	0.61	93	93	99	0.23	0.14	96
Sample 4 (N = 6000)												
M1PL	1.28	--			-0.17	0.59				0.19	--	
M2PL	1.34	0.51			-0.16	0.60	98			0.19	--	
3PL	1.56	0.51	69		-0.06	0.61	93	92		0.23	0.14	
Parms	1.44	0.45	76	96	-0.09	0.61	93	93	100	0.23	0.14	98

Data/ Model	Equating Set Items (n=30)											
	Item Discrimination				Item Difficulty					Guessing		
	Mean	SD	M2PL	3PL	Mean	SD	M1PL	M2PL	3PL	Mean	SD	3PL
Sample 1 (N = 600)												
M1PL	1.29	--			-0.11	0.68				0.19	--	
M2PL	1.47	0.53			-0.15	0.69	98			0.19	--	
3PL	1.68	0.57	83		-0.05	0.69	94	93		0.23	0.14	
Parms	1.49	0.47	79	79	-0.10	0.68	96	96	96	0.21	0.09	62
Sample 2 (N = 900)												
M1PL	1.30	--			-0.11	0.69				0.19	--	
M2PL	1.45	0.50			-0.14	0.71	97			0.19	--	
3PL	1.53	0.54	82		-0.14	0.74	96	98		0.18	0.10	
Parms	1.49	0.47	84	89	-0.10	0.68	96	97	98	0.21	0.09	61
Sample 3 (N = 1200)												
M1PL	1.28	--			-0.11	0.67				0.19	--	
M2PL	1.40	0.49			-0.14	0.72	97			0.19	--	
3PL	1.55	0.53	83		-0.07	0.69	91	94		0.24	0.13	
Parms	1.49	0.47	75	90	-0.10	0.68	96	98	96	0.21	0.09	65
Sample 4 (N = 1500)												
M1PL	1.28	--			-0.10	0.67				0.19	--	
M2PL	1.39	0.44			-0.11	0.65	98			0.19	--	
3PL	1.60	0.53	80		-0.04	0.63	96	96		0.23	0.11	
Parms	1.49	0.47	76	95	-0.10	0.68	95	97	99	0.21	0.09	83

[a] Correlations with decimals omitted are listed to the right of the means and SDs.

Table 3
Summary of Relationships between Generating Item Parameters (Parms) and Estimates (M1PL, M2PL, and 3PL) - Section 2[a]

	Operational Items (n = 38)											
	Item Discrimination				Item Difficulty					Guessing		
Data/ Model	Mean	SD	M2PL	3PL	Mean	SD	M1PL	M2PL	3PL	Mean	SD	3PL
Sample 1 (N = 2400)												
M1PL	1.21	--			0.05	0.52				0.20	--	
M2PL	1.17	0.44			-0.02	0.63	96			0.20	--	
3PL	1.34	0.42	80		0.07	0.68	93	96		0.23	0.11	
Parms	1.31	0.35	73	90	0.08	0.62	93	93	97	0.24	0.12	81
Sample 2 (N = 3600)												
M1PL	1.21	--			0.02	0.53				0.20	--	
M2PL	1.21	0.41			-0.02	0.59	97			0.20	--	
3PL	1.40	0.43	68		0.06	0.64	94	94		0.21	0.13	
Parms	1.32	0.35	74	93	0.08	0.62	93	94	98	0.24	0.12	86
Sample 3 (N = 4800)												
M1PL	1.21	--			0.03	0.54				0.20	--	
M2PL	1.18	0.41			-0.02	0.61	97			0.20	--	
3PL	1.30	0.41	78		0.05	0.68	95	95		0.23	0.12	
Parms	1.32	0.35	74	94	0.08	0.62	93	94	99	0.24	0.12	88
Sample 4 (N = 6000)												
M1PL	1.24	--			0.02	0.52				0.20	--	
M2PL	1.23	0.43			-0.02	0.58	96			0.20	--	
3PL	1.36	0.41	72		0.06	0.66	92	93		0.23	0.13	
Parms	1.32	0.35	73	95	0.08	0.62	93	94	100	0.24	0.12	96

	Equating Set Items (N=20)											
	Item Discrimination				Item Difficulty					Guessing		
Data/ Model	Mean	SD	M2PL	3PL	Mean	SD	M1PL	M2PL	3PL	Mean	SD	3PL
Sample 1 (N = 600)												
M1PL	1.21	--			0.09	0.68				0.20	--	
M2PL	1.37	0.49			0.08	0.71	98			0.20	--	
3PL	1.60	0.56	79		0.18	0.65	92	94		0.25	0.15	
Parms	1.40	0.49	73	84	0.06	0.71	96	97	94	0.21	0.07	48
Sample 2 (N = 900)												
M1PL	1.21	--			0.09	0.68				0.20	--	
M2PL	1.40	0.49			0.06	0.68	96			0.20	--	
3PL	1.60	0.57	80		0.11	0.69	93	97		0.22	0.11	
Parms	1.40	0.41	70	80	0.06	0.71	95	97	98	0.21	0.07	84
Sample 3 (N = 1200)												
M1PL	1.21	--			0.09	0.68				0.20	--	
M2PL	1.36	0.41			0.05	0.72	97			0.20	--	
3PL	1.42	0.47	79		0.05	0.73	95	98		0.19	0.10	
Parms	1.40	0.41	78	87	0.06	0.71	96	98	99	0.21	0.07	82
Sample 4 (N = 1500)												
M1PL	1.24	--			0.09	0.70				0.20	--	
M2PL	1.38	0.45			0.07	0.73	98			0.20	--	
3PL	1.38	0.46	89		0.05	0.76	97	99		0.19	0.08	
Parms	1.40	0.41	74	88	0.06	0.71	97	99	100	0.21	0.07	94

[a] Correlations with decimals omitted are listed to the right of the means and SDs.

Table 4
Summary of Relationships between Generating Item Parameters (Parms) and Estimates (M1PL, M2PL, and 3PL) - Section 3[a]

Data/ Model	Operational Items (n = 58) Item Discrimination				Item Difficulty					Guessing		
	Mean	SD	M2PL	3PL	Mean	SD	M1PL	M2PL	3PL	Mean	SD	3PL
Sample 1 (N = 2400)												
M1PL	1.34	--			0.03	0.68				0.20	--	
M2PL	1.37	0.39			0.01	0.74	99			0.20	--	
3PL	1.42	0.44	80		-0.01	0.76	97	98		0.18	0.10	
Parms	1.40	0.41	72	95	0.04	0.76	97	98	99	0.21	0.09	89
Sample 2 (N = 3600)												
M1PL	1.30	--			0.05	0.71				0.20	--	
M2PL	1.30	0.36			0.02	0.78	99			0.20	--	
3PL	1.36	0.40	82		0.03	0.80	98	98		0.20	0.09	
Parms	1.40	0.41	76	97	0.04	0.76	97	98	100	0.21	0.09	88
Sample 3 (N = 4800)												
M1PL	1.31	--			0.03	0.70				0.20	--	
M2PL	1.33	0.37			0.02	0.75	99			0.20	--	
3PL	1.42	0.41	81		0.03	0.76	98	98		0.20	0.09	
Parms	1.40	0.41	76	97	0.04	0.76	97	98	100	0.21	0.09	95
Sample 4 (N = 6000)												
M1PL	1.32	--			0.05	0.69				0.20	--	
M2PL	1.32	0.38			0.04	0.75	99			0.20	--	
3PL	1.37	0.43	78		0.04	0.80	97	98		0.20	0.10	
Parms	1.40	0.41	76	97	0.04	0.76	97	97	99	0.21	0.09	93

Data/ Model	Equating Set Items (N=30) Item Discrimination				Item Difficulty					Guessing		
	Mean	SD	M2PL	3PL	Mean	SD	M1PL	M2PL	3PL	Mean	SD	3PL
Sample 1 (N = 600)												
M1PL	1.34	--			-0.06	0.80				0.20	--	
M2PL	1.44	0.48			-0.06	0.82	98			0.20	--	
3PL	1.49	0.52	89		-0.06	0.79	96	98		0.21	0.11	
Parms	1.41	0.42	85	85	-0.06	0.77	96	98	98	0.21	0.10	68
Sample 2 (N = 900)												
M1PL	1.30	--			-0.06	0.78				0.20	--	
M2PL	1.42	0.41			-0.07	0.76	99			0.20	--	
3PL	1.45	0.42	85		-0.05	0.74	96	99		0.22	0.11	
Parms	1.41	0.42	73	77	-0.06	0.77	95	98	98	0.21	0.10	65
Sample 3 (N = 1200)												
M1PL	1.31	--			-0.06	0.78				0.20	--	
M2PL	1.39	0.41			-0.05	0.78	99			0.20	--	
3PL	1.50	0.45	85		-0.02	0.75	96	97		0.22	0.12	
Parms	1.41	0.42	86	91	-0.06	0.77	95	98	98	0.21	0.10	75
Sample 4 (N = 1500)												
M1PL	1.32	--			-0.06	0.79				0.20	--	
M2PL	1.41	0.44			-0.05	0.79	98			0.20	--	
3PL	1.34	0.41	94		-0.09	0.81	98	99		0.18	0.08	
Parms	1.41	0.42	88	90	-0.06	0.77	95	98	99	0.21	0.10	43

[a] Correlations with decimals omitted are listed to the right of the means and SDs.

Table 5
Values of Common C and Number of Operational and Equating Items
Set at Common C by LOGIST by Sample and Section*

	Section 1	Section 2	Section 3
Sample 1	.19379 (13) (7)	.18830 (11) (4)	.14103 (9) (4)
Sample 2	.18959 (11) (6)	.16061 (9) (5)	.17636 (10) (4)
Sample 3	.19167 (13) (7)	.18748 (7) (4)	.15094 (9) (4)
Sample 4	.19085 (13) (7)	.16684 (9) (5)	.14825 (10) (5)

*Number of operational and equating set items, respectively, fixed at Common C are listed in parentheses to the right of the Common C values.

Table 6
Summary of Relationships between Generating Ability Parameters and Estimates*

	Section 1					Section 2					Section 3				
Data/ Model	Mean	SD	M1PL	M2PL	3PL	Mean	SD	M1PL	M2PL	3PL	Mean	SD	M1PL	M2PL	3PL
Sample 1															
M1PL	0.29	0.68				0.51	0.72				0.28	0.73			
M2PL	0.30	0.72	99			0.53	0.79	99			0.29	0.75	100		
3PL	0.28	0.68	99	100		0.50	0.74	99	100		0.27	0.74	100	100	
Parms	0.27	0.64	94	94	95	0.47	0.71	93	93	93	0.28	0.73	96	96	96
Sample 2															
M1PL	0.29	0.69				0.48	0.72				0.33	0.75			
M2PL	0.30	0.73	99			0.49	0.76	99			0.33	0.78	100		
3PL	0.29	0.72	99	100		0.47	0.71	99	100		0.32	0.78	100	100	
Parms	0.28	0.66	94	95	95	0.47	0.70	93	93	93	0.30	0.71	96	96	96
Sample 3															
M1PL	0.31	0.68				0.49	0.72				0.29	0.74			
M2PL	0.31	0.71	99			0.50	0.78	99			0.30	0.76	100		
3PL	0.30	0.69	99	100		0.48	0.77	99	100		0.29	0.74	100	100	
Parms	0.28	0.65	94	95	95	0.46	0.70	93	93	93	0.28	0.72	96	96	96
Sample 4															
M1PL	0.29	0.68				0.47	0.70				0.31	0.74			
M2PL	0.30	0.71	99			0.48	0.75	99			0.32	0.77	100		
3PL	0.28	0.66	99	99		0.46	0.75	99	100		0.31	0.78	100	100	
Parms	0.28	0.65	95	95	95	0.46	0.70	92	92	93	0.29	0.72	96	96	96

*Correlations with decimals omitted are listed to the right of the means and SDs.

Summary of Model-Data Fit

Examples of item-ability regression plots based on the M1PL, M2PL, and 3PL models are displayed in Figures 1 through 3. Each of these figures presents plots for items 31 through 36 in Section 1 for Sample 4. Figure 1 indicates substantial misfit of the M1PL model to items 31 and 33, misfit at the lower ability levels for items 32 and 35, and adequate fit to items 34 and 36. Figure 2 indicates that the M2PL model provides adequate fit to the data for items 31, 34, and 36, but some evidence of misfit at the lower ability levels is seen for items 32, 33, and 35. Figure 3 indicates that the 3PL model provides adequate fit to all six of the featured items. The patterns in these figures are typical of the patterns seen overall, although the differences between the fit of the three models to individual items tended to be more striking in the operational items for the larger samples, where misfit at lower ability levels was easier to detect.

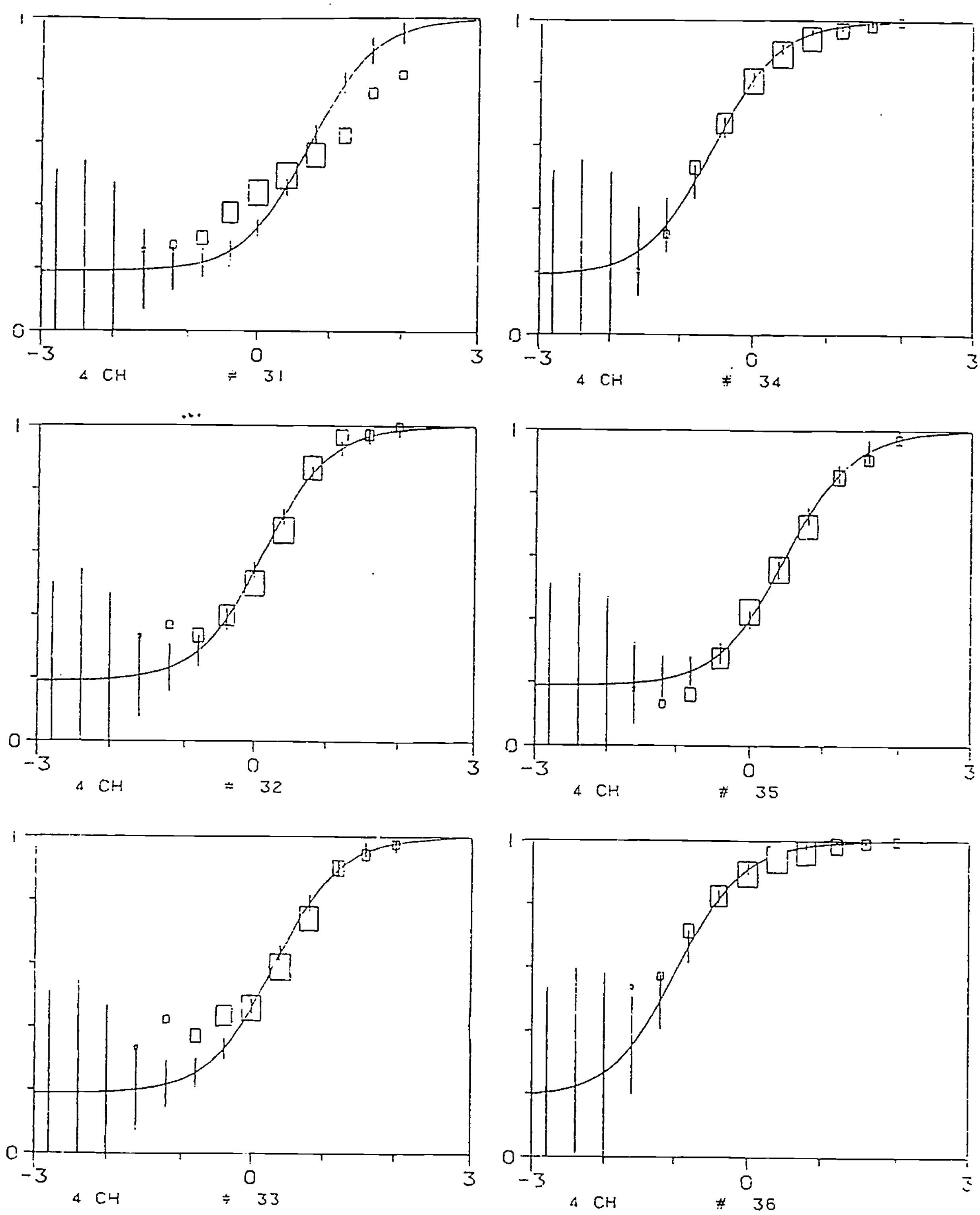

Figure 1. Selected Item-Ability Regression Plots Based on the MlPL Model - Sample 1, Section 1

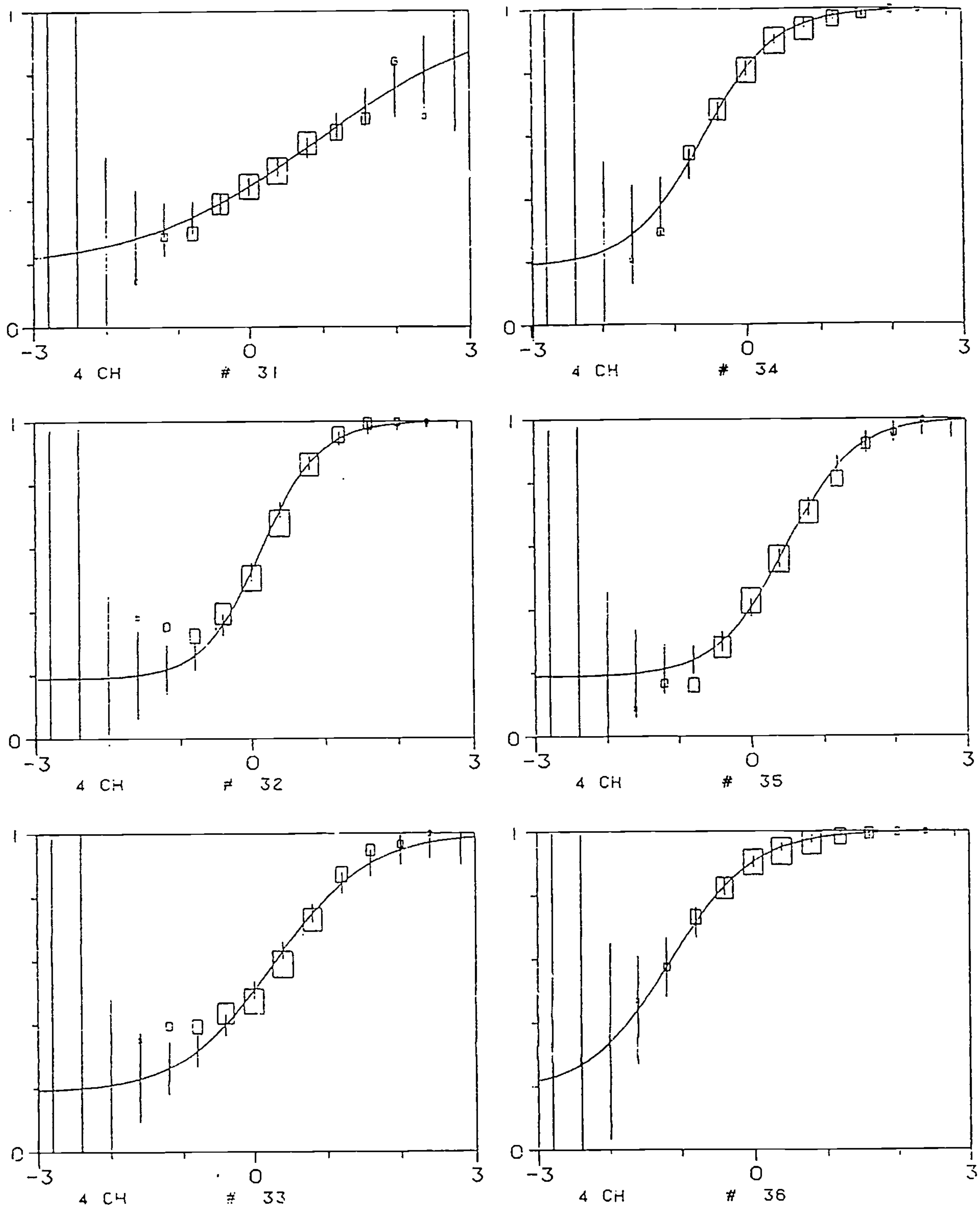

Figure 2. Selected Item-Ability Regression Plots Based on the M2PL Model - Sample 1, Section 1

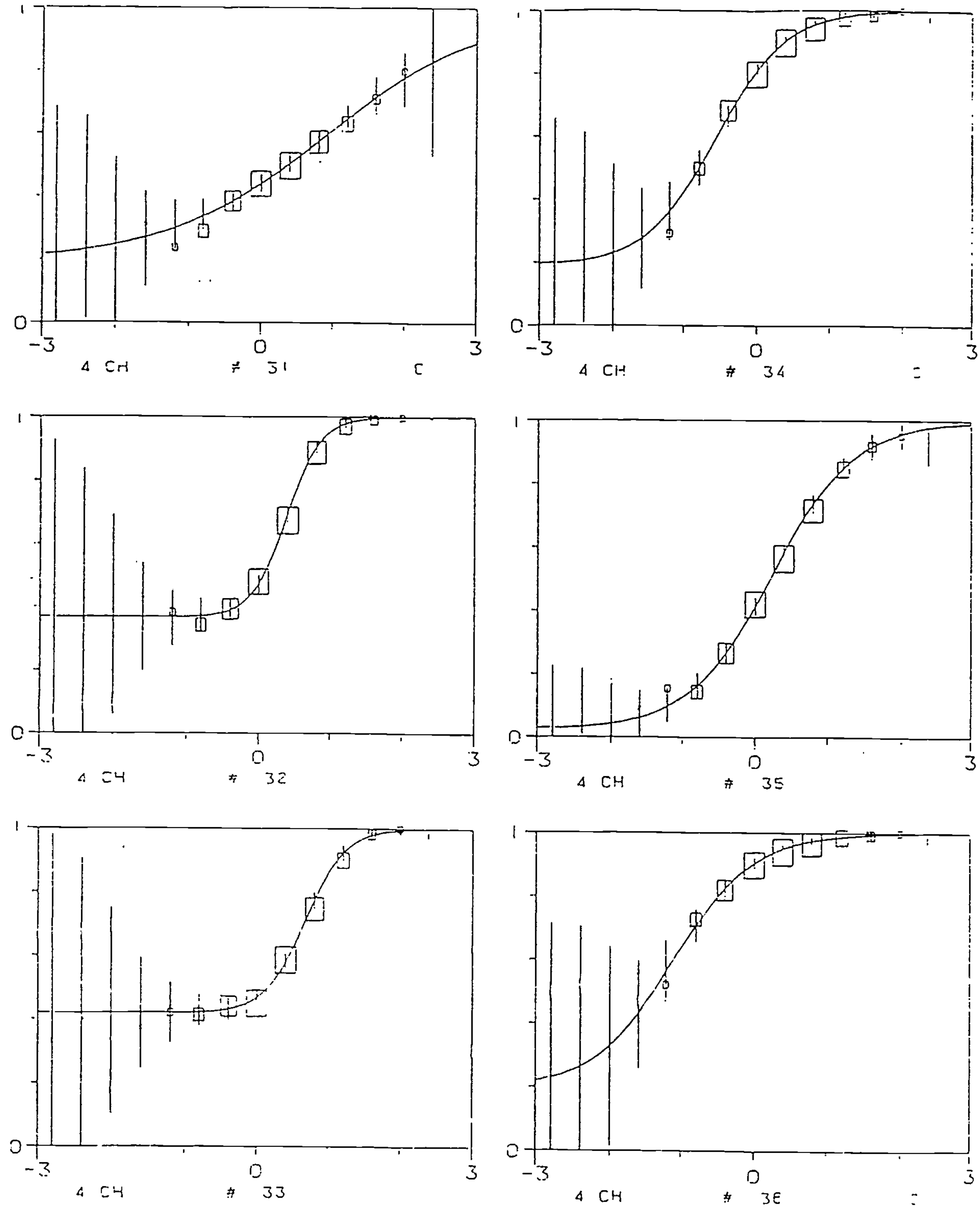

Figure 3. Selected Item-Ability Regression Plots Based on the 3PL Model - Sample 1, Section 1

Results of the residual analyses are summarized in Table 7 and include for each model, section, and sample size, the percentages of absolute standardized residuals (ASRs) between 0 and 1, between 1 and 2, between 2 and 3, and greater than 3. These categorizations are separated for the operational items and for the equating set items. In interpreting ASRs such as those in Table 7, it is generally assumed that if a given model fits the data the standardized residuals should approximate a standard normal distribution. Under this assumption, one would expect about 68% of ASRs to fall between 0 and 1, about 27% between 1 and 2, about 4% between 2 and 3, and less than 1% above 3.

Table 7
Summary of Absolute Standardized Residuals by Model, Sample, and Section

Operational Items	Section 1 (500 Residuals)				Section 2 (380 Residuals)				Section 3 (580 Residuals)			
	¦0-1¦	¦1-2¦	¦2-3¦	¦> 3¦	¦0-1¦	¦1-2¦	¦2-3¦	¦> 3¦	¦0-1¦	¦1-2¦	¦2-3¦	¦> 3¦
Sample 1												
M1PL	48.2	25.6	14.2	12.0	41.6	35.0	12.4	11.1	54.1	25.5	13.4	6.9
M2PL	64.6	27.8	5.8	1.8	68.4	24.5	6.1	1.1	68.6	25.2	4.3	1.9
3PL	72.8	24.2	2.6	0.4	71.1	26.8	1.8	0.3	73.8	24.0	2.1	0.2
Sample 2												
M1PL	41.4	26.4	14.2	18.0	41.8	31.1	15.0	12.1	45.7	28.6	15.3	10.3
M2PL	65.6	25.8	5.6	3.0	69.2	20.8	8.4	1.6	62.8	29.0	5.5	2.8
3PL	72.4	25.8	1.8	0.0	74.2	23.4	2.1	0.3	71.6	23.6	4.3	0.5
Sample 3												
M1PL	36.4	29.4	14.8	19.4	39.5	28.4	15.8	16.3	43.8	26.9	16.2	13.1
M2PL	63.0	25.0	8.0	4.0	64.7	28.2	4.2	2.9	65.7	25.0	6.7	2.6
3PL	71.2	26.0	2.6	0.2	73.9	24.5	1.1	0.5	75.7	21.4	2.9	0.0
Sample 4												
M1PL	32.8	27.6	17.4	22.2	36.6	27.1	17.1	19.2	41.9	25.2	16.6	16.4
M2PL	61.6	27.2	7.2	4.0	61.1	28.9	7.6	2.4	62.4	26.7	7.6	3.3
3PL	74.8	22.8	2.4	0.0	75.0	21.6	3.2	0.3	71.4	24.5	3.6	0.5

Equating Set Items	Section 1 (300 Residuals)				Section 2 (200 Residuals)				Section 3 (300 Residuals)			
	¦0-1¦	¦1-2¦	¦2-3¦	¦> 3¦	¦0-1¦	¦1-2¦	¦2-3¦	¦> 3¦	¦0-1¦	¦1-2¦	¦2-3¦	¦> 3¦
Sample 1												
M1PL	57.7	33.7	6.7	2.0	65.5	24.5	7.5	2.5	60.0	31.3	7.7	1.0
M2PL	66.7	28.7	4.7	0.0	73.5	20.0	6.0	0.5	72.7	24.0	2.7	0.7
3PL	72.7	25.0	2.3	0.0	74.0	20.0	5.5	0.5	77.3	19.3	2.3	1.0
Sample 2												
M1PL	53.7	35.3	8.7	2.3	60.5	27.0	10.0	2.5	60.0	29.3	9.7	1.0
M2PL	74.7	20.3	4.0	1.0	72.0	23.0	4.5	0.5	70.3	26.3	3.0	0.3
3PL	77.0	20.0	3.0	0.0	72.0	24.5	2.5	1.0	74.3	23.3	2.3	0.0
Sample 3												
M1PL	54.7	29.0	12.3	4.0	56.5	29.0	12.5	2.0	55.0	31.7	12.0	1.3
M2PL	72.7	22.7	4.3	0.3	71.0	26.0	2.5	0.5	71.0	27.0	1.7	0.3
3PL	71.3	25.7	2.7	0.3	76.0	23.5	0.5	0.0	76.7	22.7	0.7	0.0
Sample 4												
M1PL	49.3	34.7	12.3	3.7	52.0	27.5	17.5	3.0	54.7	31.3	10.7	3.3
M2PL	64.3	29.0	5.3	1.3	66.5	29.0	3.0	1.5	71.7	25.0	2.7	0.7
3PL	72.7	22.7	3.7	1.0	72.5	23.0	4.0	0.5	77.0	20.3	2.7	0.0

There are several interesting trends in Table 7. First, it is clear that fit based on the 3PL model is better than fit based on the M2PL model, and that fit based on the M2PL model is better than fit based on the M1PL model. However, what is more interesting is how the categorizations of ASRs for the three models differ according to sample size. In general, the fit of the M1PL model as reflected by the ASRs deteriorates as sample size increases. To a lesser extent, this trend is also apparent in the ASRs based on the M2PL model. Only in the case of the 3PL model is the quality of fit as indicated by the ASRs unrelated to sample size.

Because of the variety of sample sizes on which the ASRs are based, the data in Table 7 provide a clarifying framework for considering the fit of the various models. For example, for the equating set items in Section 2, 65.5% of the M1PL ASRs are between 0 and 1, 24.5% are between 1 and 2, 7.5% are be'ween 2 and 3, and 2.5% are above 3. From these percentages, it seems that even given generating parameters based on a 3PL model, by limiting data simulation to small samples one could make a convincing case for adequate fit of the M1PL model based on an analysis of residuals. In the context of the TOEFL test, if pretest items were calibrated using the M1PL model with sample sizes of 600 or lower, it is likely that there would be little evidence of model-data misfit. However, as the ASRs in Table 7 indicate, given larger sample sizes the model-data fit of the M1PL is dramatically inferior to that based on the 3PL model and even the M2PL model. Similarly, in comparing the ASRs based on the M2PL model with those based on the 3PL model, the largest discrepancies are seen in the larger sample sizes, while in the smaller sample sizes, the ASRs based on the two models reflect a similar quality of model-data fit.

Evaluation of Equating Conversions

Table 8 contains a summary of the WRMSD statistics by TOEFL test section, simulation sample, and estimation model. In this table, the contributions of the standard deviations of the converted score differences (SD Diff) and the bias to the WRMSD statistic are also given. Also included in each of the design cells of Table 8 is the WRMSD statistic expressed as a proportion of the criterion converted score standard deviations (WRMSD/Critσ). This latter statistic allows comparisons of the equating results across TOEFL sections and provides a framework for interpreting the equating errors in the simulation.

From comparisons between the statistics based on each of the three models, it is clear that the results based on the 3PL model were clearly superior to those based on the M2PL model, and that the results based on the M2PL model were superior to those based on the M1PL model. Only in Section 3 for Sample 1 was the WRMSD value based on the M2PL model lower than the WRMSD value based on the 3PL model. Only in Section 1 in Samples 1 and 4 were the WRMSD values based on the M1PL model lower than the WRMSD values based on the M2PL model. In most cases, the WRMSD values based on the 3PL model were two to three times lower than those based on the M1PL and M2PL models. In particular, for these latter models the WRMSD values for Section 2 approached and exceeded 10% of the criterion score standard deviations.

Table 8
Weighted Root Mean Square Difference Statistics by Section, Sample, and Estimation Model[*]

	Section I (n = 50)			Section II (n = 38)			Section III (n = 58)		
	M1PL	M2PL	3PL	M1PL	M2PL	3PL	M1PL	M2PL	3PL
Sample 1 (N = 2400)									
SD Diff	.3153	.3799	.0907	.8067	.6579	.1258	.5638	.3659	.4170
Bias	.2299	.1259	.0852	.4973	.4529	.2363	.1036	.0872	.0245
WRMSD	.3902	.4002	.1244	.9476	.7987	.2677	.5733	.3761	.4178
WRMSD/Critσ	.0600	.0615	.0191	.1231	.1038	.0348	.0760	.0499	.0554
Sample 2 (N = 3600)									
SD Diff	.3483	.3423	.1375	.7648	.7291	.4456	.4187	.2715	.1277
Bias	.2085	.1157	.0708	.2146	.1316	.0207	.2959	.2189	.1037
WRMSD	.4060	.3613	.1547	.7943	.7409	.4461	.5128	.3488	.1645
WRMSD/Critσ	.0624	.0555	.0238	.1031	.0961	.0579	.0691	.0470	.0222
Sample 3 (N = 4800)									
SD Diff	.3368	.3389	.0606	.7954	.7166	.2486	.4780	.3294	.2223
Bias	.2893	.2154	.1185	.3298	.2526	.0910	.1175	.0949	.0305
WRMSD	.4440	.4016	.1330	.8611	.7598	.2647	.4922	.3428	.2243
WRMSD/Critσ	.0682	.0248	.0204	.1103	.0974	.0339	.0655	.0556	.0298
Sample 4 (N = 6000)									
SD Diff	.3223	.3607	.1659	.8845	.6939	.1198	.5058	.3602	.1939
Bias	.1734	.0960	.0474	.1686	.1482	-.0313	.2272	.2107	.0962
WRMSD	.3660	.3733	.1725	.9004	.7095	.1238	.5545	.4173	.2165
WRMSD/Critσ	.0563	.0575	.0266	.1165	.0918	.0160	.0744	.0560	.0290

[*] Critσ refers to the standard deviation of the criterion scores. Critσ ranged from 6.50 to 6.51 for Section 1 samples, from 7.70 to 7.80 for Section 2 samples, and from 7.42 to 7.54 for Section 3 samples.

Somewhat surprisingly, sample size did not appear to have much influence on how well the different models reproduced the criterion score conversions. This was particularly illuminating in the case of the 3PL model. For example, for the Section 1 data sets, the lowest WRMSD value for the 3PL model was obtained in Sample 1, while the highest WRMSD value was obtained in Sample 4. This suggests that while smaller sample sizes clearly detracted from how well the individual 3PL model item parameter estimates reproduced the generating item parameters, the smaller sample sizes did not appear to affect the quality of resulting 3PL equating conversions. A priori, it was assumed that the 3PL equatings based on the smaller sample sizes would compare less favorably with the 2PL and 1PL equatings because of greater estimation error. However, as Table 8 clearly indicates, this was not the case. In considering the implications of this finding it should be mentioned that even in the case of Sample 1, the data matrix used by LOGIST was dominated by the operational items, which made up about 65% of the items calibrated and for which there were adequate numbers of responses for a 3PL model calibration. In a typical TOEFL pretest administration, the operational items make up only 30% of the total number of items calibrated, and the total data matrix is much sparser. In this case, whether or not smaller sample sizes will result in acceptable conversions with the 3PL model is less certain.

One final observation concerning Table 8 is that with one exception, all of the bias statistics were positive. For the M1PL and M2PL models, this probably occurred because a parameter for guessing was not estimated. For the 3PL model, the explanation is more complicated, and may be related to the fact that "old" item parameter estimates for equating set items were based on LOGIST "fixed b's" calibrations where the maximum item discrimination value was set at 1.50. Because historical evidence strongly suggested that this maximum value was too low, it was raised to 1.70 when the new TOEFL equating design was introduced. As a result, "new" estimates of the same items often reflect higher item discrimination values than "old" estimates, particularly after both sets of estimates have been placed on the scale of the base form. In carrying out the simulated equatings, the "new" and "old" estimates used for the equating set items reflected these differences (see footnote 1), and may have accounted for the positive biases. As previously mentioned, in Tables 2 through 4 the mean 3PL item discrimination estimate is nearly always higher than the generating mean item discrimination value. Although any systematic bias may be cause for concern, for most of the cases in Table 8 the contribution of bias to the WRMSD statistics was less than the contribution of the variation in the score differences. Furthermore, because equating with TOEFL is carried out directly to the base form, there is no opportunity for bias to accumulate over repeated administrations.

In addition to examining the WRMSD statistics, the quality of the simulated equatings based on the M1PL, M2PL, and 3PL models can be evaluated by comparing rounded raw to scaled score conversion tables with the criterion conversion tables. These tables are listed in Appendices A, B, and C for Sections 1, 2, and 3, respectively. In these tables, differences between the conversions based on all three models and the criterion conversions are seen at the extremes of the score scales, and are mostly positive. It is in these score ranges that the superiority of the 3PL model over the M1PL and M2PL models is most pronounced. At the more crucial middle points of the score range (i.e., scaled scores from 45 to 55), all three models appear to perform about equally well.

That the M1PL and M2PL models performed reasonably well in the middle of the score scale and inadequately at the extremes of the score scale is probably due to fact that the converted score scale that is being reproduced is based on a 3PL model. At the lower end of the score scale, guessing has a strong impact on conversions that are obtained with the 3PL model. At the upper end of the score scale, the positive relationships between difficulty, discrimination, and guessing that generally occur with 3PL LOGIST calibrations have an impact on conversions. In the middle of the scale, score conversions are mostly determined by the overall difficulty of the test, and less affected by differences in discrimination and guessing among individual items simulated using the 3PL model.

Conclusions

The purpose of this study was to explore the use of two alternative item response theory estimation models in the scaling and equating of TOEFL -- a modified one-parameter model (M1PL) and a modified two-parameter model (M2PL) -- and to compare item scaling and test equating results based on these two alternative models with results based on the three-parameter model (3PL) that is currently being used to scale and equate the TOEFL test. The study employed a design in which a typical TOEFL equating was simulated using artificial data. Simulated equatings were carried out for all three sections of the TOEFL using total sample sizes of 2,400, 3,600, 4,800, and 6,000. For each TOEFL section, simulated responses for operational items were generated for the complete samples, while for one-fourth of the data in each sample, responses for equating set items were also generated. The simulated equatings carried out using the M1PL, M2PL, and 3PL models were compared in terms of correlations between estimated and generating parameters, model-data fit, and concordance of simulated score conversions with conversions based on the generating parameters.

The results of the study clearly indicated that the 3PL model performed better than the M1PL and M2PL models on the basis of each of the evaluation criteria. There was also evidence that the M2PL model performed better than the M1PL model, particularly in terms of model-data fit and in the WRMSD statistics used to evaluate the simulated score conversions. The results of the study also indicated that discrepancies between score conversions based on the M1PL and M2PL model and those based on the 3PL model tended to occur at the lower and upper ends of the score scales. Finally, the results of the study for the 3PL model indicated that while correlations between item parameter estimates and generating parameters tended to be affected by sample size, neither the quality of model-data fit nor the quality of simulated equatings appeared to be sensitive to the different sample sizes used in the study. This was somewhat surprising, as it had been expected that the 3PL model would perform less had been adequately relative to the M2PL and M1PL models as the sample sizes decreased.

Some caution should be used in interpreting the results of this study, particularly because they were based on artificial data. Although the data were simulated to be as realistic as possible, it cannot be said with certainty that the same results would have been obtained if real TOEFL data had been employed. In addition, the study was somewhat limited in that the procedures used in carrying out the equatings were those that have been successfully applied with the 3PL model in operational settings. For example, the item parameter transformations were obtained using an item characteristic curve method that may not be optimal for the M1PL model. An alternative method, such as a mean and sigma transformation, might have produced better equating results with the M1PL model.

Despite these limitations, the results of this study suggest that the 3PL model should be favored over the M1PL and M2PL models for scaling and equating the TOEFL. The overall performance of these models was clearly inferior to the performance of the 3PL model, and while there was some evidence that the M1PL and M2PL models produced adequate score conversions in the most important range of the TOEFL section score scales (i.e., scaled scores of 45 to 55), lower scaled scores were often misrepresented by as many as seven points based on the M1PL model and five points based on the M2PL model. Furthermore, given the historical context of the TOEFL item banking, test development, and equating procedures, the less immediate effects of changing to a M1PL or M2PL model could actually be more severe than effects suggested by the results of this study. Although the results of this study should not be construed as completely definitive, taken together with the fact that the current TOEFL score scale is 3PL model based, it would appear that only in the context of a complete restructuring of the TOEFL test would it make sense to further consider the use of a M1PL or M2PL model. Such a restructuring would basically have to involve defining a new TOEFL item bank and a new TOEFL score scale from scratch.

With respect to the continued use of the 3PL model for equating the TOEFL test, the results of this study suggest several avenues for future investigations. For example, a study focusing specifically on calibration sample sizes would be useful. Such a study would have to focus not only on the direct effects of sample size on a single equating, but also on the indirect effects of pretest sample sizes on the quality of future equatings. A second study of interest would be an investigation of alternate algorithms and common item designs for carrying out ICC transformations as part of the TOEFL equatings. For example, in the context of the present TOEFL equating design, it would be possible to augment the external equating set items with operational items which would also have pre-test IRT statistics available. Considerations of common item designs for equating the TOEFL would also be useful in investigating the possibility of using IRT for TOEFL test assembly, which is a research project currently being considered. Finally, further research -- particularly in a simulation context -- might be useful in determining what long-term effects, if any, can be expected from the recent change from the fixed-b's equating design previously used with TOEFL, and the new design that makes use of ICC transformations.

References

Baker, F. B. (1987). Methodology Review: Item parameter estimation under the one-, two-,and three-parameter logistic models. Applied Psychological Measurement, 12, 111-141.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff.

Hambleton, R. K., & Murray, L. N. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

Hicks, M. M. (1983). True score equating by fixed b's scaling: A flexible and stable equating alternative. Applied Psychological Measurement, 7, 255-266.

Hicks, M. M. (1984). A comparative study of methods of equating TOEFL test scores. (ETS Research Report 84-20). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Marco, G. L., Wingersky, M. S., & Douglass, J. B. (1985). An evaluation of three approximate item response theory models for equating test scores. (ETS Research Report 85-46). Princeton, NJ: Educational Testing Service.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness of fit statistics. Applied Psychological Measurement, 9, 49-57.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland and D. B. Rubin (Eds.), Test Equating. Princeton, NJ: Educational Testing Service.

Skaggs, G. K., & Lissitz, R. (1986). Test equating: Relevant issues and a review of recent research. Review of Educational Research, 56, 495-529.

Stocking, M. L. (1988). Specifying optimum examinees for item parameter estimation in item response theory. (ETS Research Report 88-57). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Swaminathan, H., & Gifford, J. A. (1979). Estimation of parameters in the three-parameter latent-trait model. Laboratory of Psychometric and Evaluation Research (Report No. 90). Amherst MA: University of Massachusetts.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Wingersky, M. S., Patrick, R., & Lord, F. M. (1988). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.

Appendix A
Raw-to-Scaled Score Conversions
for TOEFL Section 1

Table A.1
Score Conversions for Section 1, Sample 1

	Crit.	M1PL		M2PL		3PL	
RS	SS	SS	Dif	SS	Dif	SS	Dif
0	23	25	+2	25	+2	24	+1
1	24	26	+2	26	+2	25	+1
2	25	27	+2	27	+2	26	+1
3	26	28	+2	28	+2	27	+1
4	27	29	+2	29	+2	27	
5	28	29	+1	29	+1	28	
6	29	30	+1	30	+1	29	
7	30	31	+1	31	+1	30	
8	30	32	+2	32	+2	31	+1
9	31	33	+2	33	+2	32	+1
10	32	35	+3	34	+2	33	+1
11	33	36	+3	36	+3	34	+1
12	35	38	+3	37	+2	36	+1
13	36	39	+3	38	+2	37	+1
14	38	39	+1	39	+1	38	
15	39	40	+1	40	+1	39	
16	40	41	+1	40		40	
17	41	42	+1	41		41	
18	42	42		42		42	
19	42	43	+1	42		42	
20	43	43		43		43	
21	44	44		44		44	
22	44	45	+1	44		44	
23	45	45		45		45	
24	46	46		45	-1	46	
25	46	46		46		46	
26	47	47		47		47	
27	47	48	+1	47		47	
28	48	48		48		48	
29	49	49		48	-1	49	
30	49	49		49		49	
31	50	50		50		50	
32	50	50		50		51	+1
33	51	51		51		51	
34	52	52		52		52	
35	52	52		52		52	
36	53	53		53		53	
37	54	54		54		54	
38	54	55	+1	54		54	
39	55	55		55		55	
40	56	56		56		56	
41	57	57		57		57	
42	57	58	+1	58	+1	57	
43	58	59	+1	59	+1	58	
44	59	60	+1	60	+1	59	
45	60	61	+1	61	+1	60	
46	61	62	+1	62	+1	61	
47	63	63		64	+1	63	
48	64	64		65	+1	64	
49	66	65	-1	67	+1	66	
50	68	68		68		68	

Table A.2
Score Conversions for Section 1, Sample 2

	Crit.	M1PL		M2PL		3PL	
RS	SS	SS	Dif	SS	Dif	SS	Dif
0	23	25	+2	25	+2	24	+1
1	24	26	+2	26	+2	25	+1
2	25	27	+2	27	+2	25	
3	26	28	+2	28	+2	26	
4	27	29	+2	29	+2	27	
5	28	29	+1	29	+1	28	
6	29	30	+1	30	+1	29	
7	30	31	+1	31	+1	30	
8	30	32	+2	32	+2	31	+1
9	31	33	+2	33	+2	32	+1
10	32	35	+3	34	+2	33	+1
11	33	36	+3	35	+2	33	
12	35	38	+3	37	+2	35	
13	36	39	+3	38	+2	36	
14	38	39	+1	39	+1	38	
15	39	40	+1	40	+1	39	
16	40	41	+1	40		40	
17	41	42	+1	41		40	-1
18	42	42		42		41	-1
19	42	43	+1	42		42	
20	43	43		43		43	
21	44	44		44		43	-1
22	44	45	+1	44		44	
23	45	45		45		45	
24	46	46		45	-1	45	-1
25	46	46		46		46	
26	47	47		47		47	
27	47	48	+1	47		47	
28	48	48		48		48	
29	49	49		48	-1	49	
30	49	49		49		49	
31	50	50		50		50	
32	50	50		50		50	
33	51	51		51		51	
34	52	52		52		52	
35	52	52		52		52	
36	53	53		53		53	
37	54	54		54		54	
38	54	55	+1	54		54	
39	55	55		55		55	
40	56	56		56		56	
41	57	57		57		57	
42	57	58	+1	58	+1	58	+1
43	58	59	+1	59	+1	58	
44	59	59		60	+1	59	
45	60	60		61	+1	60	
46	61	61		62	+1	62	+1
47	63	63		63		63	
48	64	64		65	+1	64	
49	66	65	-1	67	+1	66	
50	68	68		68		68	

Table A.3
Score Conversions for Section 1, Sample 3

	Crit.	M1PL		M2PL		3PL	
RS	SS	SS	Dif	SS	Dif	SS	Dif
0	23	25	+2	25	+2	23	
1	24	26	+2	26	+2	24	
2	25	27	+2	27	+2	25	
3	26	28	+2	28	+2	26	
4	27	29	+2	29	+2	27	
5	28	29	+1	29	+1	28	
6	29	30	+1	30	+1	29	
7	30	31	+1	31	+1	30	
8	30	32	+2	32	+2	31	+1
9	31	33	+2	33	+2	32	+1
10	32	35	+3	34	+2	32	
11	33	36	+3	36	+3	33	
12	35	38	+3	37	+2	35	
13	36	39	+3	38	+2	37	+1
14	38	39	+1	39	+1	38	
15	39	40	+1	40	+1	39	
16	40	41	+1	41	+1	40	
17	41	42	+1	41		41	
18	42	42		42		42	
19	42	43	+1	43	+1	42	
20	43	43		43		43	
21	44	44		44		44	
22	44	45	+1	44		44	
23	45	45		45		45	
24	46	46		46		46	
25	46	46		46		46	
26	47	47		47		47	
27	47	48	+1	47		48	+1
28	48	48		48		48	
29	49	49		49		49	
30	49	49		49		49	
31	50	50		50		50	
32	50	51	+1	50		51	+1
33	51	51		51		51	
34	52	52		52		52	
35	52	53	+1	52		52	
36	53	53		53		53	
37	54	54		54		54	
38	54	55	+1	54		54	
39	55	55		55		55	
40	56	56		56		56	
41	57	57		57		57	
42	57	58	+1	58	+1	58	+1
43	58	59	+1	59	+1	58	
44	59	60	+1	60	+1	59	
45	60	61	+1	61	+1	60	
46	61	62	+1	62	+1	61	
47	63	63		63		63	
48	64	64		65	+1	64	
49	66	65	-1	67	+1	66	
50	68	68		68		68	

Table A.4
Score Conversions for Section 1, Sample 4

	Crit.	M1PL		M2PL		3PL	
RS	SS	SS	Dif	SS	Dif	SS	Dif
0	23	25	+2	25	+2	23	
1	24	26	+2	26	+2	24	
2	25	27	+2	27	+2	25	
3	26	28	+2	28	+2	26	
4	27	29	+2	29	+2	27	
5	28	29	+1	29	+1	28	
6	29	30	+1	30	+1	29	
7	30	31	+1	31	+1	29	-1
8	30	32	+2	32	+2	30	
9	31	33	+2	33	+2	31	
10	32	35	+3	34	+2	32	
11	33	36	+3	36	+3	33	
12	35	37	+2	37	+2	35	
13	36	38	+2	38	+2	37	+1
14	38	39	+1	39	+1	38	
15	39	40	+1	40	+1	39	
16	40	41	+1	41	+1	40	
17	41	41		41		41	
18	42	42		42		42	
19	42	43	+1	43	+1	43	+1
20	43	43		43		43	
21	44	44		44		44	
22	44	45	+1	44		45	+1
23	45	45		45		45	
24	46	46		45	-1	46	
25	46	46		46		46	
26	47	47		47		47	
27	47	47		47		48	+1
28	48	48		48		48	
29	49	49		48	-1	49	
30	49	49		49		49	
31	50	50		50		50	
32	50	50		50		51	+1
33	51	51		51		51	
34	52	52		52		52	
35	52	52	.	52		52	
36	53	53		53		53	
37	54	54		54		54	
38	54	54		54		54	
39	55	55		55		55	
40	56	56		56		56	
41	57	57		57		57	
42	57	58	+1	58	+1	57	
43	58	59	+1	58		58	
44	59	60	+1	60	+1	59	
45	60	61	+1	61	+1	60	
46	61	62	+1	62	+1	61	
47	63	63		63		62	-1
48	64	64		65	+1	64	
49	66	65	-1	67	+1	66	
50	68	68		68		68	

Appendix B
Raw-to-Scaled Score Conversions
for TOEFL Section 2

Table B.1
Score Conversions for Section 2, Sample 1

	Crit.	M1PL		M2PL		3PL	
RS	SS	SS	Dif	SS	Dif	SS	Dif
0	20	20		20		20	
1	20	20		20		20	
2	20	21	+1	21	+1	20	
3	21	23	+2	23	+2	21	
4	22	24	+2	24	+2	22	
5	23	25	+2	25	+2	24	+1
6	24	26	+2	26	+2	25	+1
7	25	27	+2	27	+2	26	+1
8	26	31	+5	29	+3	27	+1
9	27	34	+7	31	+4	29	+2
10	30	35	+5	33	+3	32	+2
11	33	37	+4	35	+2	34	+1
12	35	38	+3	36	+1	36	+1
13	37	39	+2	38	+1	37	
14	38	40	+2	39	+1	38	
15	39	41	+2	40	+1	39	
16	40	42	+2	41	+1	41	+1
17	41	42	+1	42	+1	42	+1
18	42	43	+1	42		42	
19	43	44	+1	43		43	
20	44	45	+1	44		44	
21	45	45		45		45	
22	46	46		46		46	
23	46	47	+1	46		47	+1
24	47	48	+1	47		47	
25	48	48		48		48	
26	49	49		49		49	
27	50	50		50		50	
28	51	51		51		51	
29	52	52		52		52	
30	53	53		53		53	
31	54	54		54		54	
32	55	55		55		55	
33	56	56		56		56	
34	57	57		58	+1	57	
35	59	59		60	+1	59	
36	61	61		62	+1	61	
37	63	63		65	+2	63	
38	68	68		68		68	

Table B.2
Score Conversions for Section 2I, Sample 2

	Crit.	M1PL		M2PL		3PL	
RS	SS	SS	Dif	SS	Dif	SS	Dif
0	20	20		20		20	
1	20	20		20		20	
2	20	21	+1	21	+1	20	
3	21	23	+2	23	+2	22	+1
4	22	24	+2	24	+2	23	+1
5	23	25	+2	25	+2	24	+1
6	24	26	+2	26	+2	25	+1
7	25	27	+2	27	+2	26	+1
8	26	31	+5	30	+4	28	+2
9	27	33	+6	32	+5	31	+4
10	30	35	+5	34	+4	33	+3
11	33	36	+3	36	+3	35	+2
12	35	38	+3	37	+2	36	+1
13	37	39	+2	38	+1	38	+1
14	38	40	+2	39	+1	39	+1
15	39	40	+1	40	+1	40	+1
16	40	41	+1	41	+1	41	+1
17	41	42	+1	42	+1	42	+1
18	42	43	+1	42		42	
19	43	44	+1	43		43	
20	44	44		44		44	
21	45	45		45		45	
22	46	46		45	-1	46	
23	46	47	+1	46		46	
24	47	47		47		47	
25	48	48		48		48	
26	49	49		49		49	
27	50	50		49	-1	50	
28	51	51		50	-1	51	
29	52	52		51	-1	51	-1
30	53	53		52	-1	52	-1
31	54	54		53	-1	53	-1
32	55	55		55		54	-1
33	56	56		56		56	
34	57	57		57		57	
35	59	59		59		58	-1
36	61	61		62	+1	60	-1
37	63	63		65	+2	63	
38	68	68		68		68	

Table B.3
Score Conversions for Section 2, Sample 3

	Crit.	M1PL		M2PL		3PL	
RS	SS	SS	Dif	SS	Dif	SS	Dif
0	20	20		20		20	
1	20	20		20		20	
2	20	21	+1	21	+1	20	
3	21	23	+2	23	+2	21	
4	22	24	+2	24	+2	22	
5	23	25	+2	25	+2	24	+1
6	24	26	+2	26	+2	25	+1
7	25	27	+2	27	+2	26	+1
8	26	31	+5	29	+3	27	+1
9	27	34	+7	32	+5	29	+2
10	30	35	+5	34	+4	32	+2
11	33	37	+4	35	+2	34	+1
12	35	38	+3	36	+1	35	
13	37	39	+2	38	+1	37	
14	38	40	+2	39	+1	38	
15	39	40	+1	40	+1	39	
16	40	41	+1	41	+1	40	
17	41	42	+1	41		41	
18	42	43	+1	42		42	
19	43	44	+1	43		43	
20	44	44		44		44	
21	45	45		45		45	
22	46	46		45	-1	46	
23	46	47	+1	46		46	
24	47	47		47		47	
25	48	48		48		48	
26	49	49		49		49	
27	50	50		50		50	
28	51	51		50	-1	51	
29	52	52		51	-1	52	
30	53	53		52	-1	53	
31	54	54		54		54	
32	55	55		55		55	
33	56	56		56		56	
34	57	57		58	+1	57	
35	59	59		60	+1	59	
36	61	61		62	+1	61	
37	63	63		65	+2	64	+1
38	68	68		68		68	

Table B.4
Score Conversions for Section 2, Sample 4

	Crit.	M1PL		M2PL		3PL	
RS	SS	SS	Dif	SS	Dif	SS	Dif
0	20	20		20		20	
1	20	20		20		20	
2	20	21	+1	21	+1	20	
3	21	23	+2	23	+2	21	
4	22	24	+2	24	+2	22	
5	23	25	+2	25	+2	23	
6	24	26	+2	26	+2	24	
7	25	27	+2	27	+2	26	+1
8	26	31	+5	29	+3	27	+1
9	27	34	+7	32	+5	28	+1
10	30	35	+5	34	+4	31	+1
11	33	37	+4	36	+3	33	
12	35	38	+3	37	+2	35	
13	37	39	+2	38	+1	37	
14	38	40	+2	39	+1	38	
15	39	41	+2	40	+1	39	
16	40	41	+1	41	+1	40	
17	41	42	+1	42	+1	41	
18	42	43	+1	42		42	
19	43	44	+1	43		43	
20	44	44		44		44	
21	45	45		45		45	
22	46	46		46		46	
23	46	47	+1	46		46	
24	47	47		47		47	
25	48	48		48		48	
26	49	49		49		49	
27	50	50		49	-1	50	
28	51	51		50	-1	51	
29	52	51	-1	51	-1	52	
30	53	52	-1	52	-1	53	
31	54	53	-1	53	-1	54	
32	55	54	-1	55		55	
33	56	56		56		56	
34	57	57		57		57	
35	59	58	-1	59		59	
36	61	60	-1	61		61	
37	63	63		64	+1	63	
38	68	68		68		68	

Appendix C
Raw-to-Scaled Score Conversions
for TOEFL Section 3

Table C.1
Score Conversions for Section 3, Sample 1

	Crit.	M1PL		M2PL		3PL	
RS	SS	SS	Dif	SS	Dif	SS	Dif
0	20	20		20		20	
1	20	21	+1	21	+1	21	+1
2	21	22	+1	22	+1	22	+1
3	22	23	+1	23	+1	23	+1
4	23	23		23		24	+1
5	24	24		24		25	+1
6	24	25	+1	25	+1	25	+1
7	25	26	+1	26	+1	26	+1
8	26	27	+1	27	+1	27	+1
9	27	27		27		28	+1
10	27	28	+1	28	+1	29	+2
11	28	29	+1	29	+1	30	+2
12	29	31	+2	30	+1	31	+2
13	30	32	+2	31	+1	32	+2
14	31	33	+2	32	+1	33	+2
15	32	34	+2	34	+2	34	+2
16	33	35	+2	35	+2	35	+2
17	34	36	+2	36	+2	36	+2
18	35	37	+2	37	+2	37	+2
19	37	38	+1	38	+1	38	+1
20	38	39	+1	39	+1	39	+1
21	39	40	+1	40	+1	40	+1
22	40	41	+1	40		40	
23	41	42	+1	41		41	
24	42	42		42		42	
25	43	43		43		43	
26	43	44	+1	44	+1	44	+1
27	44	44		44		44	
28	45	45		45		45	
29	46	46		46		46	
30	46	46		46		46	
31	47	47		47		47	
32	48	48		48		48	
33	48	48		48		48	
34	49	49		49		49	
35	50	49	-1	49	-1	50	
36	50	50		50		50	
37	51	51		51		51	
38	51	51		51		51	
39	52	52		52		52	
40	52	52		52		52	
41	53	53		53		53	
42	53	53		53		53	
43	54	54		54		54	
44	55	54	-1	54	-1	54	-1
45	55	55		55		55	
46	56	56		56		55	-1
47	56	56		56		56	
48	57	57		57		57	
49	57	57		57		57	
50	58	58		58		58	
51	59	59		59		59	
52	60	60		60		59	-1
53	60	61	+1	61	+1	60	
54	61	61		61		61	
55	63	62	-1	63		62	-1
56	64	64		64		64	
57	65	65		65		65	
58	67	67		67		67	

Table C.2
Score Conversions for Section 3, Sample 2

	Crit.	M1PL		M2PL		3PL	
RS	SS	SS	Dif	SS	Dif	SS	Dif
0	20	20		20		20	
1	20	21	+1	21	+1	21	+1
2	21	22	+1	22	+1	22	+1
3	22	23	+1	23	+1	22	
4	23	23		23		23	
5	24	24		24		24	
6	24	25	+1	25	+1	25	+1
7	25	26	+1	26	+1	25	
8	26	27	+1	27	+1	26	
9	27	27		27		27	
10	27	28	+1	28	+1	28	+1
11	28	29	+1	29	+1	29	+1
12	29	31	+2	30	+1	30	+1
13	30	32	+2	31	+1	31	+1
14	31	33	+2	32	+1	32	+1
15	32	34	+2	33	+1	33	+1
16	33	35	+2	34	+1	34	+1
17	34	36	+2	35	+1	35	+1
18	35	37	+2	36	+1	36	+1
19	37	38	+1	37		37	
20	38	39	+1	38		38	
21	39	40	+1	39		39	
22	40	41	+1	40		40	
23	41	42	+1	41		41	
24	42	42		42		42	
25	43	43		43		43	
26	43	44	+1	44	+1	43	
27	44	45	+1	44		44	
28	45	45		45		45	
29	46	46		46		46	
30	46	47	+1	46		46	
31	47	47		47		47	
32	48	48		48		48	
33	48	48		48		48	
34	49	49		49		49	
35	50	50		50		50	
36	50	50		50		50	
37	51	51		51		51	
38	51	51		51		51	
39	52	52		52		52	
40	52	52		52		52	
41	53	53		53		53	
42	53	54	+1	54	+1	54	+1
43	54	54		54		54	
44	55	55		55		55	
45	55	55		55		55	
46	56	56		56		56	
47	56	57	+1	57	+1	56	
48	57	57		57		57	
49	57	58	+1	58	+1	58	+1
50	58	59	+1	59	+1	58	
51	59	59		59		59	
52	60	60		60		60	
53	60	61	+1	61	+1	61	+1
54	61	62	+1	62	+1	62	+1
55	63	63		63		63	
56	64	64		64		64	
57	65	65		66	+1	66	+1
58	67	67		67		67	

Table C.3
Score Conversions for Section 3, Sample 3

	Crit.	M1FL		M2PL		3PL	
RS	SS	SS	Dif	SS	Dif	SS	Dif
0	20	20		20		20	
1	20	21	+1	21	+1	21	+1
2	21	22	+1	22	+1	22	+1
3	22	23	+1	23	+1	22	
4	23	23		23		23	
5	24	24		24		24	
6	24	25	+1	25	+1	25	+1
7	25	26	+1	26	+1	25	
8	26	27	+1	27	+1	26	
9	27	27		27		27	
10	27	28	+1	28	+1	28	+1
11	28	29	+1	29	+1	29	+1
12	29	31	+2	30	+1	30	+1
13	30	32	+2	31	+1	31	+1
14	31	33	+2	32	+1	32	+1
15	32	34	+2	33	+1	33	+1
16	33	35	+2	34	+1	34	+1
17	34	36	+2	35	+1	35	+1
18	35	37	+2	36	+1	36	+1
19	37	38	+1	37		37	
20	38	39	+1	38		38	
21	39	40	+1	39		39	
22	40	41	+1	40		40	
23	41	41		41		41	
24	42	42		42		42	
25	43	43		43		43	
26	43	44	+1	44	+1	44	+1
27	44	44		44		44	
28	45	45		45		45	
29	46	46		46		46	
30	46	46		46		47	+1
31	47	47		47		47	
32	48	48		48		48	
33	48	48		48		48	
34	49	49		49		49	
35	50	49	-1	49	-1	50	
36	50	50		50		50	
37	51	51		51		51	
38	51	51		51		51	
39	52	52		52		52	
40	52	52		52		52	
41	53	53		53		53	
42	53	53		53		53	
43	54	54		54		54	
44	55	54	-1	55		54	-1
45	55	55		55		55	
46	56	56		56		56	
47	56	56		56		56	
48	57	57		57		57	
49	57	58	+1	58	+1	57	
50	58	58		58		58	
51	59	59		59		59	
52	60	60		60		60	
53	60	61	+1	61	+1	60	
54	61	62	+1	62	+1	61	
55	63	63		63		63	
56	64	64		64		64	
57	65	65		66	+1	65	
58	67	67		67		67	

Table C.4
Score Conversions for Section 3, Sample 4

	Crit.	M1PL		M2PL		3PL	
RS	SS	SS	Dif	SS	Dif	SS	Dif
0	20	20		20		20	
1	20	21	+1	21	+1	21	+1
2	21	22	+1	22	+1	22	+1
3	22	23	+1	23	+1	22	
4	23	23		23		23	
5	24	24		24		24	
6	24	25	+1	25	+1	25	+1
7	25	26	+1	26	+1	26	+1
8	26	27	+1	27	+1	26	
9	27	27		27		27	
10	27	28	+1	28	+1	28	+1
11	28	29	+1	29	+1	29	+1
12	29	31	+2	30	+1	30	+1
13	30	32	+2	31	+1	31	+1
14	31	33	+2	33	+2	32	+1
15	32	34	+2	34	+2	33	+1
16	33	35	+2	35	+2	34	+1
17	34	36	+2	36	+2	35	+1
18	35	37	+2	37	+2	36	+1
19	37	38	+1	38	+1	37	
20	38	39	+1	39	+1	38	
21	39	40	+1	40	+1	39	
22	40	41	+1	40		40	
23	41	42	+1	41		41	
24	42	42		42		42	
25	43	43		43		43	
26	43	44	+1	44	+1	43	
27	44	45	+1	44		44	
28	45	45		45		45	
29	46	46		46		46	
30	46	47	+1	46		46	
31	47	47		47		47	
32	48	48		48		48	
33	48	48		48		48	
34	49	49		49		49	
35	50	50		50		50	
36	50	50		50		50	
37	51	51		51		51	
38	51	51		51		51	
39	52	52		52		52	
40	52	52		52		52	
41	53	53		53		53	
42	53	53		53		54	+1
43	54	54		54		54	
44	55	55		55		55	
45	55	55		55		55	
46	56	56		56		56	
47	56	56		56		56	
48	57	57		57		57	
49	57	58	+1	58	+1	58	+1
50	58	58		58		58	
51	59	59		59		59	
52	60	60		60		60	
53	60	61	+1	61	+1	61	+1
54	61	62	+1	62	+1	62	+1
55	63	63		63		63	
56	64	64		64		64	
57	65	65		66	+1	66	+1
58	67	67		67		67	

TOEFL is a program of
Educational Testing Service
Princeton, New Jersey, USA

57906-06414 • D21P.5 • 275704 • Printed in U.S A